統計的手法による分野非依存のテキスト分割

内山　将夫[†]　井佐原　均[†]

複数のトピックからなる文章を，それぞれのトピックに切り分けることをテキスト分割と呼ぶ．テキスト分割は，情報検索や要約のための基本技術として有用である．本稿では，分割確率最大化という観点からテキスト分割を定式化した．その定式化の特色の一つは，テキスト内の単語しか，確率推定に利用しないことである．そのため，提案手法は，任意の分野のテキストに対して適用できる．提案手法の有効性は二つの実験により確認された．まず，実験1では，公開データに対して提案手法を適用することにより，提案手法の分割精度が従来手法の分割精度よりも優れていることが示された．次に，実験2では，長い文書の元々の章や節の構造と提案手法による分割結果とを比較した結果，厳密な一致のみを正解とする場合，章には0.37, 節には0.34の割合で一致し，±1行のずれを許容する場合，章には0.49, 節には0.51の割合で一致した．これらのことは，提案手法が，テキスト分割に対して有効であることを示している．

キーワード:　テキスト分割，統計的手法

A Statistical Approach to Domain Independent Text Segmentation

Masao Utiyama[†] and Hitoshi Isahara[†]

A text is usually composed of multiple topics. Segmenting such a text into coherent topics is useful both for information retrieval and for automatic text summarization. This paper proposes a statistical method that selects the segmentation of the highest probability among possible segmentations as the best segmentation of the given text. Since the method estimates probabilities of segmentations from the given text, it does not need training data. Therefore, it can be applied to any text in any domain. The effectiveness of the method was confirmed through two experiments. The first experiment evaluated the accuracy of the method by using publicly available data. The experimental results showed that the accuracy of the proposed method is at least as good as that of a state-of-the-art text segmentation system. The second experiment compared the segmentations done by our method with those of original segments in relatively long documents. When we compared our system's segmentations with chapters in the documents, the accuracy was 0.37 on the condition that we regarded only exact matches as correct matches. If we regarded ± 1 line differences as correct then the accuracy was 0.49. When we compared our system's segmentations with sections, the accuracies were 0.34 and 0.51, respectively. These results show that our method is effective for domain independent text segmentation.

KeyWords:　*text segmentation, statistical approach*

† 通信総合研究所, Communications Research Laboratory

1　はじめに

ある程度の長さの文章は，一般的に，複数のトピックからなる．そのような文章を切り分けて，それぞれの切り分けた部分が一つのトピックになるようにすることを，テキスト分割と呼ぶ．

テキスト分割は，情報検索や要約などにおいて重要である．まず，情報検索においては，文書全体ではなく，ユーザの検索要求を満す部分 (トピック) だけを検索した方が効果的である (Hearst and Plaunt 1993; Salton, Singhal, Buckley, and Mitra 1996; 望月 奥村 2000)．また，要約においては，長い文書をトピックに分ければ，それぞれのトピックごとに要約を作成することにより，文書全体の要約を作成できるし，重要なトピックだけを選んで要約を作成することもできる (Kan, Klavans, and McKeown 1998; Nakao 2000)．

これらの目的のために，多くの手法が研究されている (Kozima 1993; Hearst 1994; Okumura and Honda 1994; Salton et al. 1996; Yaari 1997; Kan et al. 1998; Choi 2000; Nakao 2000; 望月・奥村 2000, など)．これらの手法の主な共通点は，これらの手法が，分割対象のテキスト (および辞書やシソーラス) しか分割に利用しないことである．たとえば，(Hearst 1994) は，テキスト内の単語分布の類似度しか分割に利用しない．言い換えれば，これらの手法は，その手法をテキスト分割に使用するにあたって，訓練データを必要としない．

そのため，これらの手法は，訓練データの存在する分野に限られることなく，どんな分野の文章でも分割対象とすることができる．この点は重要である．なぜなら，情報検索や要約が対象とする文書は，分野を限定しない文書であるので，そのような文書に対応するためには，分野を限定しないテキスト分割の手法が必要であるからである．

本稿で述べる手法も，これらの従来手法と同様に，訓練データを利用せずに，テキスト内の単語分布のみを利用してテキストを分割する．我々が，訓練データを利用しないテキスト分割手法を採用した理由は，我々が，テキスト分割の結果を利用して，長い文書を要約したり，講演のディクテーション結果を要約することを目的としているからである．そのためには，分野を限定しない (訓練データを利用しない) テキスト分割の方法が必要であるからである．

本稿で述べる手法は，テキストの分割確率が最大となるような分割を選択するというものである．このようなアプローチは，分野を限定しないテキスト分割としては，新しいアプローチである．なお，従来の研究で，分野を限定しないテキスト分割の研究では，主に，語彙的な結束性を利用してテキストを分割している．その例としては，意味ネットワーク上での活性伝播に基づく結束性を利用するもの (Kozima 1993) や，単語分布の類似度 (コサイン) を結束性としたもの (Hearst 1994) や，単語の繰り返し状況に基づいて結束性を計るもの (Reynar 1994) や，文間の類似度としてコサインを直接使うのではなくコサインの順位を結束性の指標とするもの (Choi 2000) などがある．

なお，テキスト分割の方法としては，訓練データを利用しない (分野を限定しない) 方法の他

に，訓練データを利用する方法もある．そのような方法の応用としては，複数ニュースを個々のニュースに分割するものがある (Allan, Carbonell, Doddington, Yamron, and Yang 1998)．この場合には，分野が明確であり，また，訓練データも多量にあるので，訓練データを利用したシステムにより，ニュースの境界を推定し分割する手法が主流である (Yamron, Carp, Lowe, and van Mulbregt 1998; Beeferman, Berger, and Lafferty 1999, など)．しかし，そのような方法は，訓練データが利用できない分野については適用できないので，我々の目的である，テキスト分割の結果を利用して，長い文書を要約したり，講演のディクテーション結果を要約するためのテキスト分割手法としては適さない．

以下，2章では，テキスト分割のための統計的モデルを述べ，3章で，最大確率の分割を選択するアルゴリズムを述べる．4章では，まず，我々の手法を公開データに基づいて評価することにより，我々の手法が他の手法よりも優れた分割精度を持つことを示し，次に，我々の手法を長い文書に適用した場合の分割精度を述べる．5章は考察，6章は結論である．

2　テキスト分割のための統計的モデル

本章では，テキストの分割結果の確率を定義し，それを用いて最大確率であるような分割を定義する．そして，次章で，最大確率であるような分割を選ぶアルゴリズムを示す．

本章では，テキスト W が与えられたときに，その任意の分割 S に対して，条件付き確率 $\Pr(S|W)$ を定義する．$\Pr(S|W)$ は，テキスト W を条件とする分割 S の条件付き確率であるので，この値が最大の分割 $\hat{S}$ を選ぶことにより，W が指定された場合の最大確率の分割 $\hat{S}$ を選ぶことができる．このような分割 $\hat{S}$ は，テキスト W の本来の分割の推定として適当であると考えられる．

まず，n 個の延べ単語からなるテキスト $W = w_1 w_2 \ldots w_n$ が与えられたとき，m 個の区間からなる分割 $S = S_1 S_2 \ldots S_m$ の確率 $\Pr(S|W)$ は，

$$\Pr(S|W) = \frac{\Pr(W|S)\Pr(S)}{\Pr(W)} \tag{1}$$

である．ここで，$\Pr(W|S)$ と $\Pr(S)$ については，詳しくは，以下で定義するが，$\Pr(W|S)$ は，分割 S が与えられたときに，テキスト W が生起する確率であり，$\Pr(S)$ は，分割 S の確率である．また，$\Pr(W)$ は，テキスト W の確率であるが，これは，W が与えられているときには，定数であるから，最大確率の分割を求める際には無視できる．よって，最大確率の分割 $\hat{S}$ は，

$$\hat{S} = \arg\max_S \Pr(W|S)\Pr(S) \tag{2}$$

である．以下では，$\hat{S}$ を最適分割と呼ぶことにする．

次に，2.1節で $\Pr(W|S)$ を定義し，2.2節で $\Pr(S)$ を定義する．

2.1 $\Pr(W|S)$ の定義

区間 S_i に n_i 個の延べ単語があるとして，S_i 中の j 番目の単語を w_j^i とし，$W_i = w_1^i w_2^i \ldots w_{n_i}^i$ とする．つまり，S_i と W_i とを一対一に対応させる．このようにすると，$n = \sum_{i=1}^{m} n_i$，$W = W_1 W_2 \ldots W_m$ である．

このとき，ある区間に属する単語列は，その他の区間には独立に生起するとし，更に，同一区間に属する単語も，区間が与えられているという条件下では確率的に独立であるとすると，

$$\begin{aligned}\Pr(W|S) &= \Pr(W_1 W_2 \ldots W_m|S) \\ &= \prod_{i=1}^{m} \Pr(W_i|S) \\ &= \prod_{i=1}^{m} \Pr(W_i|S_i) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \Pr(w_j^i|S_i) \end{aligned} \tag{3}$$

である．この式の，2行目と3行目は，「ある区間に属する単語列は他の区間とは独立に生起する」という仮定から変形でき，最後の行は，「同一区間に属する単語は，その区間が与えられているという条件では，その他の単語と確率的に独立である」という仮定から変形できる．また，$\Pr(W_i|S_i)$ は，区間 S_i で単語列 W_i が生起する確率であり，$\Pr(w_j^i|S_i)$ は，区間 S_i で単語 w_j^i が生起する確率である．

次に，W 中における異なり単語の数を k，W_i において w_j^i と同じ単語[1]の数を $f_i(w_j^i)$ とし，

$$\Pr(w_j^i|S_i) \equiv \frac{f_i(w_j^i) + 1}{n_i + k} \tag{4}$$

と定義する．ここで，(4) 式は，ラプラス推定 (Laplace's law) と呼ばれる確率推定式 (Manning and Schütze 1999) である[2]．なお，$f_i(w_j^i)$ は，厳密には，次式で定義される．

$$f_i(w_j^i) \equiv g(w_j^i|w_1^i w_2^i \ldots w_{n_i}^i) \tag{5}$$

$$g(w_j^i|w_1^i w_2^i \ldots w_{n_i}^i) \equiv \sum_{k=1}^{n_i} \delta(w_k^i, w_j^i). \tag{6}$$

ただし，δ については，単語 a と単語 b とが同じとき $\delta(a,b) = 1$，そうでないとき，$\delta(a,b) = 0$ である．

1 トークンとしては異なるがタイプとしては同じということである．たとえば，$W_i = aababab$ のとき，W_i 中には，同一タイプである a が異なるトークンとして4回出現する．よって，$f_i(a) = 4$ である．

2 確率推定のその他の方法の一つとして最尤推定がある．最尤推定の場合には，$\Pr(w_j^i|S_i) \equiv \frac{f_i(w_j^i)}{n_i}$ と推定できるが，最尤推定の推定精度は，一般に，観測事象の数 (この場合には n_i) が大きくないと良くないことが知られており，観測事象の数が少ないときには，何らかのスムージングが必要である．ラプラス推定は，そのようなスムージング方法の一つである．たとえば，最尤推定によると，ある区間 S_i に一回も出現しない単語の確率は，$\frac{0}{n_i} = 0$ と推定されるが，ラプラス推定では，一回も出現しない単語についても，$\frac{0+1}{n_i+k} > 0$ の確率が割当てられる．

2.2 $\Pr(S)$ の定義

分割 S に対する事前確率 $\Pr(S)$ の定義に関しては，任意性が大きい．たとえば，同じ区間数からなる分割であっても，各区間の長さが揃っている分割の方を，長さが不揃いの分割よりも優先したい場合には，長さが揃っている分割の事前確率を大きくすべきである．しかし，ここでの我々の仮定は，そのような優先すべき分割がないというものであるので，そのような優先すべき分割を前提としないような事前確率を設定しなくてはならない．

我々は，事前確率 $\Pr(S)$ の設定において，(Stolcke and Omohundro 1994) と同様に，記述長にもとづく事前確率を与えることにした．以下では，分割確率最大化と MDL(Minimum Description Length, 最小記述長) 原理 (山西 韓 1992) との関係について極く簡単に述べ，その後で，記述長に基づいた $\Pr(S)$ の設定について述べる．なお，MDL 原理とは，「与えられたデータを，モデル自身の記述も含めて最も短く符号化できるような確率モデルが最良のモデルである」と主張するものである．

分割確率最大化と MDL 原理との関係

我々は，確率最大であるような分割を得るために，(2) 式の右辺にある

$$\Pr(W|S)\Pr(S) \tag{7}$$

を最大化しようとしているが，これは，

$$-\log\Pr(W|S) - \log\Pr(S) \tag{8}$$

を最小化しようとしていることと等価である．このことは，MDL 原理の観点からは，分割 S が与えられたときのテキスト W の記述長 $-\log\Pr(W|S)$ と，分割 S の記述長 $-\log\Pr(S)$ との和を最小化しようとしていることになる．なぜなら，一般に，ある事象 X の確率が $\Pr(X)$ のときには，X を記述 (符号化) するために必要な最小記述長は $-\log\Pr(X)$ であるからである．ただし，ここで，log の底は 2 である．

このように，最小記述長であるような分割を選択することと，最大確率であるような分割を選択することとは同等である．

記述長に基づく事前確率

以上の議論の逆から言えば，分割 S に対して，適当な記述長 $l(S)$ を割当てた場合には，その記述長を利用して，

$$\Pr(S) = 2^{-l(S)} \tag{9}$$

と定義できる[3]．なぜなら，$l(S) = -\log \Pr(S)$ であるからである．つまり，分割 S の記述長を求めることにより，その事前確率を求めることができる．よって，以下では，分割 S の記述長を求めることにより，その事前確率を求めることにする．

ここで，我々に，既に，分割対象のテキストが与えられているとすると，分割 S を指定するために必要な情報は，各区間の長さ，$n_1, n_2, \ldots, n_m$ のみである．たとえば，我々に，既に，$W = abcdefghi$ という長さが 9 のテキストが与えられていると仮定すると，そのテキストの分割を指定するためには，たとえば，2,3,3,1 という 4 つの数字からなる数字列を指定すればよい．そうすれば，W を $W = [ab][cde][fgh][i]$ のように 4 分割できる．

つまり，我々は，m 個の区間からなる分割を指定 (記述) するためには，m 個の数字を指定すれば良い．次に，これらの個々の数字は，1 以上 n 以下の n 個のうちの一つであることに注意すると，これらの個々の数字は，$1/n$ の確率で選択されると考えることができるので，$\log n$ の記述長で記述できる．よって，m 個の数字を記述するためには，$m \log n$ の記述長があれば良い．以上より，$l(S) = m \log n$ と計算できる[4]．

そのため，$\Pr(S)$ は

$$\Pr(S) \equiv n^{-m} \tag{10}$$

と定義できる．

一般的にいって，$\Pr(S)$ の値は，分割数が小さいほど大きな値を取る．一方，$\Pr(W|S)$ の値は，分割数が大きいほど大きな値を取る．そのため，もし，分割を推定するのに，$\Pr(W|S)$ だけを利用した場合には，推定される分割結果は，分割数が大きい分割，すなわち，細かすぎる区間からなる．それに対して，$\Pr(S)$ と $\Pr(W|S)$ の両方を利用した場合には，両者のバランスの取れた分割が得られる．

3　最適分割を選択するアルゴリズム

本章では，分割 S のコスト $C(S)$ を，

$$C(S) \equiv -\log \Pr(W|S) \Pr(S) \tag{11}$$

と定義し，このコストが最小となる分割 $\hat{S} = \arg\min_S C(S)$ を選択することにより，最大確率である分割 $\hat{S}$ を選択する．ここで，$C(S)$ は以下のように展開できる．

$$\begin{aligned} C(S) &= -\log \Pr(W|S) \Pr(S) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \Pr(w_j^i | S_i) + m \log n \end{aligned}$$

3 このように定義した場合，全ての分割 S に対する $\Pr(S)$ の和は 1 以下となる．つまり，$\sum_S \Pr(S) \leq 1$ となる (山西・韓 1992)．

4 このような m 個の数字を記述するための記述長には，いくつかの変種がある．それらについては，(Stolcke and Omohundro 1994) を参照のこと．

$$= \sum_{i=1}^{m} c(w_1^i w_2^i \ldots w_{n_i}^i | n, k). \tag{12}$$

ただし，

$$\begin{aligned} c(w_1^i w_2^i \ldots w_{n_i}^i | n, k) &\equiv \sum_{j=1}^{n_i} \log \frac{n_i + k}{f_i(w_j^i) + 1} + \log n \\ &= \sum_{j=1}^{\#(w_1^i w_2^i \ldots w_{n_i}^i)} \log \frac{\#(w_1^i w_2^i \ldots w_{n_i}^i) + k}{g(w_j^i | w_1^i w_2^i \ldots w_{n_i}^i) + 1} + \log n. \end{aligned} \tag{13}$$

ここで，$\#(\cdots)$ は，その引数である単語列の長さ (延べ単語数) である．なお，(13) 式を，その最終行において，n_i や f_i を使わないで定義する理由は，次節で述べるアルゴリズムにおいて，(13) 式を使うときの便宜を考えてのことである．

次に，最小コスト分割 (最大確率分割) である $\hat{S}$ を求めるアルゴリズムを示す．

3.1 最小コスト分割を求めるアルゴリズム

まず，用語を定義する．延べ語数 n のテキスト $W = w_1 w_2 \ldots w_n$ において，i 番目の分割候補点 g_i とは，単語 w_i と w_{i+1} の間を言う．ただし，g_0 は w_1 の直前，g_n は w_n の直後である．このとき，分割候補点は $g_0, g_1, \ldots, g_n$ の $n+1$ 個ある．また，分割候補点の集合をノード集合とするグラフを考えるとき，$e_{ij} (0 \leq i < j \leq n)$ は g_i から g_j への有向辺である．このように定義されたグラフの例を，図 1 に示す．

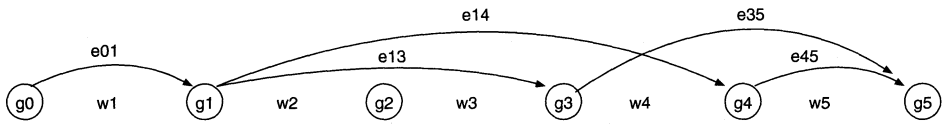

図 1　分割候補点をノードとするグラフ

このとき，e_{ij} は，単語列 $w_{i+1} w_{i+2} \ldots w_j$ をカバーするという．e_{ij} は，テキストの，ある一区間 $w_{i+1} w_{i+2} \ldots w_j$ を表現している．そのため，e_{ij} のコスト c_{ij} を，(13) 式を利用することにより，次式で定義する．

$$c_{ij} = c(w_{i+1} w_{i+2} \ldots w_j | n, k) \tag{14}$$

ただし，k は，W 中の異なり単語数である．

以上の準備の下で，最小コストを与える最適分割を求める手順は以下の通りである．

Step 1.　有向辺 e_{ij} のコスト c_{ij} を (14) 式により計算する．$(0 \leq i < j \leq n)$.

Step 2.　g_0 から g_n までの最小コストパスを求める．

ここで，Step 2 を効率的に解くアルゴリズムは良く知られている[5]．なお，Step 2 は，全ての可能なパスの中での大域的な最小コストパスを求めるものであるが，そうする代りに，パスの長さを指定した最小コストパスを求めることもできる．そのようにして求められた最小コストパスは，区間数を指定した場合の最適分割に対応している．

このようにして求めた最小コストパスについて，その各辺にカバーされる単語列を，それぞれ一つの区間とすると，それは最適分割である．たとえば，図 1で，$e_{01}e_{13}e_{35}$ が最小コストパスであるとすると，最適分割は，$[w_1][w_2w_3][w_4w_5]$ である．

なお，実際にテキストを分割するときには，全ての分割候補点を考慮するのではなく，たとえば，文と文の間でのみテキストを分割したい場合がある．その場合には，分割位置として許される分割候補点の間にのみ有向辺を張るようにすれば良い．そして，そのグラフ上での最小コストパスを探索すれば良い．次節では，我々は，文間のみでテキストが分割されると仮定して議論している．

3.2　最小コスト分割よりも細かい分割をする際の問題点と解決策

前節で述べたように，グラフの最小コストパスを求めることにより，大域的な最小コストパスによる分割だけでなく，区間数を指定した最小コストパスによる分割を求めることもできる．しかし，予備実験の結果から，指定された区間数が，もし，大域的な最小コストパスにより求められる分割の区間数よりも，ある程度以上に大きいときには，1 文や 2 文からなる小さい区間が生じやすいことが判った．このことは，大域的な最小コストパスによる分割のみが必要な場合，あるいは，大域的な最小コストパスによる分割よりも大雑把な分割が必要な場合には問題ではない．しかし，大域的な最小コストパスによる分割よりも細かい分割が必要なときには，問題である．

そこで，我々は，大域的な最小コストパスよりも細かい分割が必要なときには，まず，文章全体を大域的な最小コストパスにより分割し，そのあとで，各々の区間を，その区間を一つの文章として，再帰的に分割することにした[6]．

5 Step 2 は日本語の形態素解析においてコスト最小解 (確率最大の解) を探索するアルゴリズムと同一 (実際はより簡単) であるので，DP(Dynamic Programming) を用いて効率的に解くことができる (Nagata 1994)．また，本稿で述べた手法を実装したプログラムが第 1 著者より入手できる．なお，DP を用いてテキストを分割する研究としては，(Ponte and Croft 1997; Heinonen 1998) がある．

6 予備実験の結果から，我々の方法は，1000 文を越すような長い文章が与えられたときでも，その大域的な最小コストパスによる分割の区間数は 10 から 20 程度であることが分かった．それと逆に，新聞の社説のような比較的短いものについても，4 から 6 程度の区間数の分割が最適分割となる場合が多い．この性質は，我々がテキスト分割の結果を要約に利用しようとしているという観点からは望ましいものである．なぜならば，要約においては，文章の長さに関わらず，それを適当に少ないトピックにまとめる必要があるので，分割の結果得られる区間数は，文章の長さに，それほど影響されない方が望ましいからである．なお，このように，我々の手法において，分割の区間数が文章の長さに必ずしも比例しない理由は，(12) 式の，$m \log n$ における $\log n$ が，長い文章ほど大きくなるので，長い文章においては，短かい文章よりも分割が抑制されやすいからである．

このとき，各々の区間を分割するときの区間数は，その区間の長さの，全体の長さにおける割合に比例するようにした．たとえば，1000文からなる文章を20区間に分割したいときに，大域的な最小コストパスにより，200,400,300,100文からなる四つの区間が得られたときには，それぞれの区間を，4,8,6,2だけの区間に分割する．なお，分割数に余りがでるときには，その他の区間よりも大きい区間を，他よりも一つだけ余分に分割するようにした．たとえば，上述の文章を22に分割したいときには，それぞれを，4,8+1,6+1,2だけの区間に分割する．

このようにすれば，1文や2文からなる小さい区間が生じにくいようにすることができる[7]．このプロセスは，必要な分割数が得られるまで再帰的に実行できるが，4.2節で必要な，100程度までの分割数に対しては，1回だけの再帰で十分であった．なお，再帰的な分割の効果については，4.2節で確認する．

4　実験

本章では，まず，実験1で，我々の手法を公開データに基づいて評価することにより，我々の手法が他の手法よりも優れた分割精度を持つことを示し，次に，実験2で，我々の手法を長い文書に適用した場合の分割精度を述べる．

二つの実験の本稿全体における位置付けは以下の通りである．まず，実験1の目的は，提案手法と従来手法とを比較することにより，提案手法が，従来手法よりも，テキストを精度良く分割できることを示すことにある．そのため，もし，提案手法と従来手法とを比較したいだけならば，実験1のみで十分である．したがって，本稿の主要な目的である，提案手法の他の手法に対する優位性を示すためには，実験1だけで十分である[8]．

しかし，我々の最終的な目的は，テキスト分割の結果を，長い文書の要約 (Nakao 2000) や

7 再帰的分割により細かい分割も妥当にできる定性的な理由は以下の通りである：まず，(12) 式のコスト関数は，$C(S) = \log \frac{1}{\Pr(W|S)} + \log \frac{1}{\Pr(S)}$ である．$C(S)$ の第1項をデータのコストと呼び，第2項をモデルのコストと呼ぶことにする．一般に，データのコストは，分割が細かいほど小さくなり，モデルのコストは分割が細かいほど大きくなる．そして，最小コスト解は，これらのバランスがとれたところとなる．ところが分割を最小コスト解よりも細かくすると，モデルのコストがデータのコストよりも大幅に大きくなるため，分割の決定においてデータのコストが反映されにくくなり，妥当な分割が得られなくなる．一方，再帰的に分割したときには，再帰的な分割の対象となる各区間においては，(12) 式の m も n も再帰的な分割をする前と比べて小さい値となるため，モデルのコストが小さくなる．そのため，モデルのコストとデータのコストのバランスが取れ，妥当な分割が得られやすくなる．以上をまとめると，

再帰前	データのコスト	$\ll$	モデルのコスト
		$\Rightarrow$	データのコストが分割に反映されない
		$\Rightarrow$	データを無視した妥当でない分割となる
再帰後	データのコスト	$\sim$	モデルのコスト
		$\Rightarrow$	データのコストが分割に反映される
		$\Rightarrow$	データを考慮した妥当な分割となる．

8 もちろん，この言明は，実験1で用いたデータによる分割結果の精度が，どれほど現実のテキストの分割結果の精度を反映しているかによる．我々は，この分割結果の精度が，そのまま現実のテキストにおける分割結果の精度となることはないとしても，この分割結果で明かになる，テキスト分割システムの精度の順位は，現実のテキストにおいても反映されると考えている．また，現在，我々が入手可能なデータの中では，実験1に用いたデータが，最も包括的に従来手法を網羅しているため，各種手法を比較するテストベッドとしては妥当であると考える．

講演のディクテーション結果の要約に使うことであるので，その目的のために，提案手法が，どれほど役に立つかを調べたい．そのために，実験2においては，提案手法による分割が，どの程度，元の文書の章や節と一致するかを調べることにより，提案手法の，長い文書を要約するときへの応用の可能性を把握することを目的とする．そのため，実験2の位置付けは，今後我々の手法を実際の応用へと適用させるための前段階と考えている．我々は，将来的には，何らかのタスクに基づいて提案手法を評価することを考えている．

4.1 実験1：公開データによる評価

実験1で用いたデータは，(Choi 2000) により，各種のテキスト分割手法を比較するために用いられたデータである[9]．Choi は，彼の提案手法 $C99$ と，TextTiling(Hearst 1994), DotPlot(Reynar 1998), Segmenter(Kan et al. 1998) を比較し，$C99$ では，他の手法と比較し，誤り確率が半減されたと述べている．ただし，誤り確率とは，テキストを構成する単位 (単語，文，パラグラフ等) について，任意に選んだ r 単位だけ離れた二つの単位が誤って分割される確率のことである．ここで，r は，正しい分割における各区間の長さの平均の半分が良いとされている (Beeferman et al. 1999)．なお，実験1における r の単位は単語である． また，誤り確率が低いほど精度は良い．

この実験データは，700個のテキストからなり，個々のテキストは，10個のテキスト断片を連結したものである．そして，それぞれのテキスト断片は，Brown Corpus からランダムサンプリングされたテキストの最初の s 行である．個々のテキストは，s により特徴付けられる．表1には，実験データの諸元を示す．

表 1 実験データの諸元 (Choi 2000)

s の範囲	$3-11$	$3-5$	$6-8$	$9-11$
テキスト数	400	100	100	100

各テキストは，Choi のパッケージにあるライブラリを利用した stemmer により正規化され，その正規化されたテキストが提案手法により分割された．ただし，分割可能な位置は，(Choi 2000) と同様に，文間のみである．その後，分割されたテキストの誤り確率は，Choi のパッケージにある評価プログラムにより計算された．

その評価結果を表2と表3に示す．これらの表において，$U00$ は，提案手法において，大域的な最小コスト分割を求めたときの評価結果であり，$U00_{(b)}$ は，提案手法において，区間数を

9 `http://www.cs.man.ac.uk/~choif/software/C99-1.2-release.tgz` より入手可能である．このパッケージを展開したときにできる `naacl00Exp/data/{1,2,3}/{3-11,3-5,6-8,9-11}/*` を実験に用いた．

10 に指定[10]したときの評価結果である．また，$C99$ は，Choi のアルゴリズムによる最適分割の評価結果であり，$C99_{(b)}$ は，Choi のアルゴリズムにおいて区間数を 10 に指定した場合の評価結果である[11]．また，二つの表において，** は，比較対象であるアルゴリズムの誤り確率が t 検定により，有意水準 1%で有意差があることを示す．なお，「3 – 11」などの列の数字は，それに該当するテキストにおける誤り確率の平均であり，「全体」は，全部のテキストについての誤り確率の平均である．

表 2　分割数をプログラムが決めた場合の誤り確率の比較

	3 – 11	3 – 5	6 – 8	9 – 11	全体
$U00$	11%**	13%**	6%**	6%**	10%**
$C99$	13%	18%	10%	10%	13%

表 3　分割数が指定された場合の誤り確率の比較

	3 – 11	3 – 5	6 – 8	9 – 11	全体
$U00_{(b)}$	10%**	9%	7%**	5%**	9%**
$C99_{(b)}$	12%	11%	10%	9%	11%

これらの表から，提案手法が，$C99$ あるいは $C99_{(b)}$ と，同等あるいは，より精度良くテキストを分割できると言える．そして，$C99$ あるいは $C99_{(b)}$ は，分野非依存のテキスト分割手法のなかでは，その他の従来手法よりも精度良くテキストを分割できるので，我々の提案手法が，従来手法よりも精度良くテキストを分割できることが言える．

4.2　実験 2：長い文書の章や節との一致度による評価

実験 2 では，比較的長い文章を分割し，その分割結果と元々の章や節による分割とを比較することにより，提案手法を評価した．

実験に用いたデータは，文部省年報[12]である．我々がこのデータを用いた理由は，それが公開されているということに加えて，SGML でタグ付けされているため，付録に示す簡単な Perl スクリプトにより章 (chapter) や節 (section) を切り出せるためである[13]．

10 この際には，3.2節で述べた再帰的分割は適用していない．

11 表 3の $C99_{(b)}$ の行にある数値は，(Choi 2000) の Table 6 のものと若干異なる．その理由は，元々の数値は 500 のサンプルテキストに基づいたものであるのに対して，この表のものは，700 のサンプルに基づいて我々が再実験した結果だからである (Choi, personal communication)．なお，(Choi 2000) で使われた 500 サンプルにおける $C99_{(b)}$ の誤り確率は以下の表のものである．

	3 – 11	3 – 5	6 – 8	9 – 11	全体
$C99_{(b)}$	12%	12%	9%	9%	12%

12 `http://wwwwp.mext.go.jp/download.html` よりダウンロードできる．

13 この Perl スクリプトにより切り出せないものもある．実験に用いたものは，このスクリプトにより処理可能なものであ

表 4　文部省年報の諸元

	章の数	節の数	ページ数
昭和６０年度	13	63	62
昭和６２年度	22	96	109
昭和６３年度	13	65	52
平成元年度	13	64	54
平成２年度	13	64	55

表 4には，実験に用いた文部省年報の諸元を示す．表で，章や節の数は，元のファイルでの章や節の数を数えたものであるが，ページ数は，我々が，元テキストをポストスクリプトファイルに変換して数えたものであるので，一応の目安と考えておくのが良い．

表 4に示す文部省年報には，以下の前処理が加えられた．まず，付録のスクリプトを用いて，章や節を切り出した結果から，分割位置を示す記号を除いたテキストを得た．次に，そのテキストに対して，ChaSen version 2.25(松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 2001) を適用し，その結果から，ChaSen の品詞体系における「名詞」「未知語」「記号-アルファベット」「接頭詞」に該当するもののみを抽出し，提案手法への入力とした．ただし，名詞のうちで，その下位分類が「数」「代名詞」「非自立」「特殊」「接続詞的」「動詞非自立的」に該当するものは除いた．また，平仮名だけからなる形態素も除いた．なお，このときの分割可能な位置はスクリプトの出力結果の各行の終りである．これは，段落の間で分割していることに相当する．

表 5，表 6，表 7には，このようにして得られたテキストを，分割対象の章や節の数を指定して，分割したときの精度を示す．ここで，表のタイトルに付記している再帰的分割とは，3.2節で述べた再帰的方法により分割した場合を示し，非再帰的分割とは，再帰的分割をしていない場合を示す．また，精度とは，

$$\frac{\text{元の章や節と一致した分割位置の数}}{\text{章や節の分割位置の総数}}$$

である．ただし，章や節の数を n とすると，分割位置の数は，$n-1$ である．なお，本実験で分割数を指定している理由は，本実験の目的が，指定された粒度の分割をどの程度の精度で実現できるかを調べることにあるからである．粒度を指定した分割は，長い文書から，必要に応じた長さの要約を得るときに重要である (Nakao 2000)．

これらの表において,「±０の精度」とは，システムによる分割位置が，元文書の分割位置と正確に同一な場合を一致としたときの精度である．また,「±１の精度」とは，正確に同一な場合に加えて，前後１行のずれまでを許容して一致としたときの精度である．なお，それぞれの

る．

表 5　分割結果と章の区切れとの対応 (非再帰的分割)

	章の数	± 1 の精度	± 0 の精度
昭和 6 0 年度	13	0.42 (0.019)	0.33 (0.006)
昭和 6 2 年度	22	0.52 (0.021)	0.29 (0.007)
昭和 6 3 年度	13	0.50 (0.021)	0.42 (0.007)
平成元年度	13	0.50 (0.020)	0.42 (0.007)
平成 2 年度	13	0.50 (0.020)	0.42 (0.007)
平均		0.49 (0.020)	0.37 (0.007)

表 6　分割結果と節の区切れとの対応 (非再帰的分割)

	節の数	± 1 の精度	± 0 の精度
昭和 6 0 年度	63	0.29 (0.10)	0.13 (0.033)
昭和 6 2 年度	96	0.17 (0.10)	0.07 (0.032)
昭和 6 3 年度	65	0.34 (0.11)	0.16 (0.038)
平成元年度	64	0.37 (0.11)	0.19 (0.036)
平成 2 年度	64	0.38 (0.11)	0.18 (0.036)
平均		0.31 (0.10)	0.14 (0.035)

表 7　分割結果と節の区切れとの対応 (再帰的分割)

	節の数	± 1 の精度	± 0 の精度
昭和 6 0 年度	63	0.50 (0.10)	0.31 (0.033)
昭和 6 2 年度	96	0.45 (0.10)	0.32 (0.032)
昭和 6 3 年度	65	0.48 (0.11)	0.28 (0.038)
平成元年度	64	0.56 (0.11)	0.38 (0.036)
平成 2 年度	64	0.57 (0.11)	0.40 (0.036)
平均		0.51 (0.10)	0.34 (0.035)

精度を示す列において，括弧内の数値は，精度のベースライン

$$\frac{\text{テキストにおいて一致と判定する許容範囲のサイズの合計}}{\text{テキストのサイズ}} \tag{15}$$

である (仲尾 1999)[14]．ただし，本実験の場合には，サイズは行数でカウントする．

まず，表 5における，章の数を分割数としたときの分割精度を見る．表 5では，± 1 の精度の平均が 0.49 であり，± 0 の精度の平均が 0.37 である．ここで，(Reynar 1999) では，英文テキスト 4 文書について，彼の手法による分割結果が，平均 0.25 の精度で章の区切れと一致することを述べていて，(Nakao 2000) では，ベースラインが 0.005～0.01 のとき，F 値[15]が，0.31～0.39 である．これらの結果は，± 0 の精度に対応するが，テキストが違うため，直接比較する

14 (仲尾 1999) では，(15) 式を再現率のベースラインとしているが，本実験の場合には，分割数を指定しているので，再現率と精度が一致する．

15 これは，我々の精度に相当する．なお，F 値 ($=\frac{2\times\text{再現率}\times\text{適合率}}{\text{再現率}+\text{適合率}}$) は，(Nakao 2000) の Table 3 から計算で求めた．

ことは不可能である．しかし，数値だけを比較するならば，我々の方法は，章の分割に関しては，これらの方法と比べて，少なくとも同等程度に章の区切れを再現していると考える．

次に，節の数を指定したときの分割精度を表6と表7に示す．また，表8には，最小コスト解による分割数と章や節の数とを示す．

表6は，再帰をせずに，分割数を指定して分割したときの精度を示している．このとき，±1の精度の平均が0.31であり，±0の精度の平均が0.14である．一方，再帰的分割をしたときには，表7に示すように，±1の精度の平均が0.51であり，±0の精度の平均が0.34である．これから分かるように，最小コスト解よりも粒度の細かい分割が必要なときには，再帰的分割をした方が精度良く分割ができる．なお，(Nakao 2000)では，ベースラインが0.035のときにF値が0.29であるので，再帰的分割による方法は，(Nakao 2000)と比べて，少なくとも同等程度に節の区切れを再現していると考える．

表 8　最小コスト解による分割数と章や節の数

	最小コスト解による分割数	章の数	節の数
昭和６０年度	13	13	63
昭和６２年度	12	22	96
昭和６３年度	11	13	65
平成元年度	12	13	64
平成２年度	12	13	64

5　考察と今後の課題

提案手法は，分割確率最大化という観点からテキスト分割を定式化した．これに類似の手法として，訓練データを利用したテキスト分割では，(Yamron et al. 1998)が隠れマルコフモデルに基づいて，複数ニュースを個々のニュースに分割しているが，訓練データを利用しないテキスト分割では，類似の研究はない．また，(Yamron et al. 1998)についても，彼等は，テキストの分割確率を直接扱っているのではなく，各単語を生起させるようなトピックを単語毎に求め，同一トピックの単語が連続する部分を同一トピックとする，という間接的アプローチをとっている．そのため，彼等のアプローチでは，たとえば，トピックの平均の長さなどを直接取り込むことが難しい．一方，我々のアプローチでは，このことは素直に表現できる．たとえば，(Ponte and Croft 1997)と同様に，トピックの長さxが，平均長μ，標準偏差σの正規分布$N(x|\mu,\sigma)$に従うと仮定すると，単純な拡張としては，(13)式を，$\alpha+\beta+\gamma=1$として，以

下のようにすれば，トピックの長さが平均と同じくなるような分割が優先される．

$$c(w_1^i w_2^i \ldots w_{n_i}^i | n, k, \mu, \sigma, \alpha, \beta, \gamma) = \alpha \sum_{j=1}^{\#(w_1^i w_2^i \ldots w_{n_i}^i)} \log \frac{\#(w_1^i w_2^i \ldots w_{n_i}^i) + k}{g(w_j^i | w_1^i w_2^i \ldots w_{n_i}^i) + 1} + \beta \log n + \gamma \log \frac{1}{N(\#(w_1^i w_2^i \ldots w_{n_i}^i) | \mu, \sigma)}.$$

更に，彼等の手法と我々の手法との大きな違いは，彼等が単語の確率を訓練データから推定しているのに対して，我々は，単語の確率を分割対象のテキストから推定している点である．なお，訓練データが利用可能な場合に，彼等の手法と我々の手法とを比較することは興味深いであろう．その場合には，上式で示したような，トピックの長さをコスト関数として取り込むことや，種々の手がかり表現をコスト関数に取り込むことも検討したい．

次に，提案手法のテキスト分割における特徴としては，3.2 節で述べたように，長い文章でも短かい文章でも，分割数が，大幅には変動しないというものがある．これは，短かい文章は，細かい粒度で分割し，長い文章は大雑把な粒度で分割するということである．この性質は，我々がテキスト分割をする目的が要約のため，という観点からは適した性質である．なぜなら，要約では，文章の長さに関わらず，それを適当に少ないトピックにまとめる必要があるので，分割の結果得られる区間数は，文章の長さに，それほど影響されない方が望ましいからである．しかし，応用によっては，任意に指定した粒度の分割が望ましい場合もあると考えられる．そのために，我々は，本稿では，大域的な最小コスト解よりも細かい分割が必要な場合には，再帰的な分割を適用し，それは有効ではあったが，より有効な分割方法を考えることは今後の課題としたい．そのための見込みのある方法の一つは，(仲尾 1999) で提案されているように，分割したい粒度に応じて窓の大きさを設定し，その窓内を一つの文章としてテキストを分割することである．

最後に，提案手法によると，テキストの分割の結果として，テキストの各区間における単語の確率 (密度) が自然に求まる．このような密度は，重要単語の抽出 (Bookstein, Klein, and Raita 1995) や，重要説明箇所の特定 (黒橋, 白木, 長尾 1997) に有用であることが知られている．提案手法を，このようなアプリケーションに対して適用することも興味深い．

6　おわりに

我々は，本稿において，分割確率最大化という観点から，テキスト内の情報のみを用いて，テキストを分割する手法を提案した．提案手法は，従来の手法と比べて，同等以上の精度でテキストを分割することができた．このことは提案手法がテキストの分割に有用であることを示している．

我々は，今後，実際の応用におけるテキスト分割の有効性を調べることを考えている．

付録

章や節の切り出しに用いた Perl スクリプト

```
#
# perl npaa-div.pl (chapter|section) < file.sgm
#
# ファイルの第1部(part)の chapter または section に相当する部分を
# 抜き出して，区切り(================)を入れるプログラム.
#

$type = shift;          # type is either 'chapter' or 'section'
while(<>){
    if(m&<part>&i){
        while(<>){
            last if m&</part>&i;
            if(m&<$type&i){
                print "================\n";
                while(<>){
                    last if m&</$type>&i;
                    unless(m&^<&i){
                        s/&.+?;//g;
                        s/\r//g;
                        print;
                    }
                }
                redo if m&<$type&i;
            }
        }
        last;
    }
}
print "================\n";
```

参考文献

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). "Topic Detection and Tracking Pilot Study Final Report." In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop.*

Beeferman, D., Berger, A., and Lafferty, J. (1999). "Statistical Models for Text Segmentation." *Machine Learning*, **34** (1-3), 177–210.

Bookstein, A., Klein, S. T., and Raita, T. (1995). "Detecting Content-bearing Words by Serial Clustering – Extended Abstract." In *Proc. of SIGIR '95*, pp. 319–327.

Choi, F. Y. Y. (2000). "Advances in domain independent linear text segmentation." In *Proc. of NAACL-2000.*

Hearst, M. A. (1994). "Multi-Paragraph Segmentation of Expository Text." In *Proc. of ACL'94.*

Hearst, M. A. and Plaunt, C. (1993). "Subtopic Structuring for Full-Length Document Access." In *Proc. of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–68.

Heinonen, O. (1998). "Optimal Multi-Paragraph Text Segmentation by Dynamic Programming." In *Proc. of COLING-ACL'98.*

Kan, M.-Y., Klavans, J. L., and McKeown, K. R. (1998). "Linear Segmentation and Segment Significance." In *Proc. of WVLC-6*, pp. 197–205.

Kozima, H. (1993). "Text Segmentation Based on Similarity between Words." In *Proc. of ACL'93.*

黒橋禎夫, 白木伸征, 長尾眞 (1997). "出現密度分布を用いた語の重要説明箇所の特定." 情報処理学会誌, **38** (4), 845–854.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press.

望月源 奥村学 (2000). "語彙的連鎖に基づく要約の情報検索タスクを用いた評価." 自然言語処理, **7** (4), 63–77.

Nagata, M. (1994). "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm." In *Proc. of COLING'94*, pp. 201–207.

仲尾由雄 (1999). "語彙的結束性に基づく話題の階層構成の認定." 自然言語処理, **6** (6), 83–112.

Nakao, Y. (2000). "An Algorithm for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection." In *Proc. of ACL'2000*, pp. 302–309.

Okumura, M. and Honda, T. (1994). "Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion." In *Proc. of COLING-94.*

Ponte, J. M. and Croft, W. B. (1997). "Text Segmentation by Topic." In *Proc. of the First European Conference on Research and Advanced Technology for Digital Libraries*, pp. 120–129.

Reynar, J. C. (1994). "An Automatic Method of Finding Topic Boundaries." In *Proc. of ACL-94.*

Reynar, J. C. (1998). *Topic segmentation: Algorithms and applications.* Ph.D. thesis, Computer and Information Science, University of Pennsylvania.

Reynar, J. C. (1999). "Statistical Models for Topic Segmentation." In *Proc. of ACL-99*, pp. 357–364.

Salton, G., Singhal, A., Buckley, C., and Mitra, M. (1996). "Automatic Text Decomposition Using Text Segments and Text Themes." In *Proc. of Hypertext'96.*

Stolcke, A. and Omohundro, S. M. (1994). "Best-first Model Merging for Hidden

Markov Model Induction." Technical Report TR-94-003, ICSI, Berkeley, CA. ftp://ftp.icsi.berkeley.edu/pub/techreports/1994/tr-94-003.ps.gz.

Yaari, Y. (1997). "Segmentation of Expository Texts by Hierarchical Agglomerative Clustering." In *Proc. of the Recent Advances in Natural Language Processing.*

山西健司 韓太舜 (1992). "MDL 入門：情報理論の立場から." 人工知能学会誌, **7** (3).

Yamron, J. P., Carp, I., Lowe, S., and van Mulbregt, P. (1998). "A Hidden Markov Model Approach to Text Segmentation and Event Tracking." In *Proc. of ICASSP-98.*

松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 (2001). "形態素解析システム 『茶筌』 version 2.2.5 使用説明書." 奈良先端科学技術大学院大学 松本研究室.

略歴

内山 将夫: 筑波大学第三学群情報学類卒業 (1992). 筑波大学大学院工学研究科博士課程修了 (1997). 博士 (工学). 信州大学工学部電気電子工学科助手 (1997). 郵政省通信総合研究所非常勤職員 (1999). 独立行政法人通信総合研究所任期付き研究員 (2001). 言語処理学会, 情報処理学会, ACL, 人工知能学会, 日本音響学会, 各会員.

井佐原 均: 1978 年京都大学工学部電気工学第二学科卒業. 1980 年同大学院修士課程修了. 博士 (工学). 同年通商産業省電子技術総合研究所入所. 1995 年郵政省通信総合研究所. 現在、独立行政法人通信総合研究所けいはんな情報通信融合研究センター自然言語グループリーダー. 自然言語処理, 機械翻訳の研究に従事. 言語処理学会, 情報処理学会, 人工知能学会, 日本認知科学会, ACL, 各会員.

(2000 年 12 月 22 日 受付)
(2001 年 5 月 12 日 再受付)
(2001 年 6 月 29 日 採録)